Evaluating Generative AI to Extract Qualitative Data from Peer-Reviewed Documents

Scott Spillias

scott.spillias@csiro.au

CSIRO Environment https://orcid.org/0000-0002-1310-5202

Katherine Ollerhead

University of Tasmania https://orcid.org/0000-0002-7084-5493

Matthew Andreotta

CSIRO Environment https://orcid.org/0000-0002-2996-6031

Ruby Annand-Jones

CSIRO Environment https://orcid.org/0000-0002-4616-9333

Fabio Boschetti

CSIRO Environment https://orcid.org/0000-0001-8999-6913

Joseph Duggan

Australia National University https://orcid.org/0000-0002-4146-1570

Denis Karcher

Australia National University https://orcid.org/0000-0001-8006-1943

Cecile Paris

CSIRO Data61 https://orcid.org/0000-0003-3816-0176

Rebecca Shellock

University of New South Wales https://orcid.org/0000-0002-0434-3516

Rowan Trebilco

CSIRO Environment https://orcid.org/0000-0001-9712-8016

Research Article

Keywords: Artificial Intelligence, systematic review, large language models, scientific publication, natural-language processing, future of science

Posted Date: August 26th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4922498/v1

Additional Declarations: The authors declare no competing interests.

Abstract

Uptake of AI tools in knowledge production processes is rapidly growing. Here, we explore the ability of generative AI tools to reliably extract qualitative data from peer-reviewed documents. Specifically, we evaluate the capacity of multiple AI tools to analyse literature and extract relevant information for a systematic literature review, comparing the results to those of human reviewers. We address how well AI tools can discern the presence of relevant contextual data, whether the outputs of AI tools are comparable to human extractions, and whether the difficulty of question influences the performance of the extraction. While the AI tools we tested (GPT4-Turbo and Elicit) were not reliable in discerning the presence or absence of contextual data, at least one of the AI tools consistently returned responses that were on par with human reviewers. These results highlight the utility of AI tools in the extraction phase of evidence synthesis for supporting human-led reviews and underscore the ongoing need for human oversight.

1. Introduction

Evidence-informed decision-making entails identifying, appraising and mobilising the best available evidence for the design and implementation of policy, programs and interventions (1). It is increasingly expected that policy, planning and management decisions are made using the best available evidence, and this is essential for navigating the contemporary environmental challenges that threaten human well-being and prosperity (2). However, there are various barriers to the use of environmental evidence in decision-making. The most common barriers relate to: (i) the nature of the evidence itself (i.e., its accessibility, relevance, applicability and quality), (ii) the capacity and resources of decision-makers to identify, access and read the evidence and (iii) the relationships and links between scientists and decision-makers (3, 4). One practical way of addressing these issues is the use of evidence syntheses. Evidence syntheses use robust, transparent and comprehensive methods to combine information from multiple studies and sources of evidence investigating the same topic to comprehensively understand their findings (5). They include systematic reviews, systematic maps, scoping reviews (or evidence maps) and rapid evidence assessments (or rapid reviews). Comprehensive evidence syntheses allow for a thorough overview of the research question by evaluating and summarising existing literature, and can identify gaps in knowledge, support evidence-based decision-making, and inform future studies (6, 7). The rigour provided by these methods is important for ensuring that decision-making is informed by the full range of available data when designing management interventions (8–10).

Compared to traditional literature reviews (i.e., narrative reviews), evidence syntheses can be more time-consuming and resource intensive, but they often have higher quality and strength through the application of three key principles: rigour, transparency, and replicability (3, 11). However, the total number of papers in the published literature continues to surge exponentially and the time and resources required to apply more stringent and systematic methods are increasing correspondingly, making it difficult for a single individual to have comprehensive coverage of all the literature for anything beyond very narrow or nascent topics. In 2022 alone, an estimated 5.14 million academic articles were published, and the growth rate of newly published articles is still increasing (12). Therefore, multiple reviewers are often required when creating evidence syntheses and even with multiple reviewers it is still often an extremely time-consuming task (13), and important information may be incidentally omitted or overlooked due to human error. Furthermore, inherent reviewer subjectivity and blind spots (3) novel, less biassed, and quicker methods for evidence synthesis.

One potential solution to this problem is the use of artificial intelligence (AI) to identify relevant articles and extract the desired information (14). The capability of generative AI is accelerating extraordinarily rapidly, with new advancements and technologies constantly being released (15). Previous research has shown that AI can be beneficial for systematic evidence syntheses, especially during the initial paper screening process (16–21). Historically, a range of algorithms have been used to extract entities or metadata from either abstracts or full-texts in the literature (22–24). Our previous research has shown that general-use Large Language Models like ChatGPT, which can be easily applied to novel domains without extensive training or data curation by the user, can improve the reliability of a screen for literature (19). There is mounting evidence that AI tools can be used to reliably extract basic entities or quantitative data from text (21, 24, 25). To our knowledge, whilst LLMs have been used to extract discrete entity data from abstracts (26) or discrete score values from clinical notes (27), they have not yet been evaluated on their ability to answer nuanced contextual questions from the full-texts of scientific papers.

To address this gap, we evaluated the capacity of three AI implementations to extract relevant contextual information from peer-reviewed articles and compare it to that of trained human experts. To do so, we used a case study of benefits and barriers of community-based fisheries management in Pacific island states, building upon our previous research that used this case study to evaluate the ability of ChatGPT to search peer-reviewed literature and screen papers for relevance (19). For this example research topic, we posed 11 contextual questions to ask of a set of peer-reviewed papers identified in Spillias et al., 2023, aimed at generating a qualitative data set that could be used to synthesise the current state of evidence-based knowledge in this field. We used three AI implementations, two based on GPT4-Turbo and one based on the proprietary data extraction tool at Elicit.com, and one team of human experts (extractors) extract information from 33 papers (See S1 for included papers), and then used a separate team of human experts evaluate the quality of AI responses across three criteria (evaluators; see Methods). With this procedure (Fig. 1), we aimed to answer the following research questions:

1. How well can AI tools discern the presence or absence of relevant contextual data?
2. How does the quality of AI extraction outputs compare to those done by human researchers?
3. Does the difficulty or type of question influence the performance of the extraction?

2. Results

Discerning the Existence of Contextual Data

The human reviewers analysed 33 papers (see Supplement 1) according to 11 extraction questions and recorded the presence of contextual data for 74% of the 363 paper-question pairs (e.g., for a given study, a response was recorded for a question such as 'Which country was the study conducted in?') (Fig. 2). For the other 26% paper-question pairs, the reviewers determined that there was an absence of relevant data or context for the paper-question pair (The outputs from the human and AI extractions are available at (https://github.com/s-spillias/AI_Extraction). The presence/absence results for the three AI implementations (GPT4x1,GPT4x3, Elicit) had very low agreement (inter-rater reliability; Cohen's kappa) with the human reviewers' extraction (kappa < 0.10). Elicit's agreement was especially low (kappa < 0.0), because it returned responses for all but one paper-question pair (> 99% presence), which resulted in a high rate of false positives (27%), but a near-zero rate of false negatives. The single run of GPT4-Turbo (GPT4x1) agreed with the presence/absence determination of the human reviewers in 63% of the questions and had a false positive rate of 15% and a false negative rate of 21%. The triple run of GPT4-Turbo (GPT4x3) performed slightly better, with agreement with 66% of the human reviewers' paper-question pairs. Compared to GPT4x1, GPT4x3 also reduced the number of false negatives from 21–15%, but increased the false positive rate from 15 to 19%.

Quality of AI Extractions

In assessing the quality of AI extractions, both human and AI evaluators rated AI responses on a three-point scale: -1 representing 'poor', 0 for 'fair', and 1 as 'good', across three criteria: (i) how the overall Response and Context compared to the human standard; (ii) how well the Context responded to the Question; (iii) how well the Response reflected the Context (Fig. 3). An extraction score above 0 was considered the threshold for acceptable quality. We employed a one sample t-test to examine the qualitative differences between the AI extractions and the human-established acceptable quality threshold. Specifically, we used a two-sided alternative hypothesis to determine whether the AI's extraction quality was statistically significantly different from the score of 0, which represents our benchmark for 'fair' quality. The results for the first criterion (AI Response compared to Human response) showed that GPT4x1's mean extraction quality was significantly lower than the threshold ($t = -4.00$, $p < 0.001$), suggesting that its performance was not up to the acceptable standard. Conversely, Elicit exhibited a mean quality score significantly higher than the threshold ($t = 6.91$, $p < 0.001$), indicating a generally acceptable level of performance. GPT4x3's quality was not statistically different from the threshold. Additionally, we included an AI evaluator (GPT-4-Turbo, run 5 times per paper-question pair) in the assessment process to parallel the human evaluation (See Supplemental Figure S3). Although the AI evaluator tended to rate the responses more favourably than human evaluators, a linear mixed-effects model, assessed using ANOVA, showed that this difference was not statistically significant ($p = 0.097$), and broadly followed the same trends as the human evaluators. The second criterion (Context to Question) reflected similar trends to the first, with Elicit outperforming the GPT implementations. All three implementations satisfactorily provided meaningful Responses based on the provided Contexts.

We also analysed the potential influence of the extractors themselves on the assessed quality using the same t-test and found no significant effect ($p = 0.537$). This suggests that the variation in extraction quality is attributable to the AI systems rather than the individual human evaluators. Both human evaluators and AI evaluators consistently rated Elicit as providing the highest quality responses, followed by GPT4x3 (Figure S1).

To further explore the differences between the AI implementations, we utilised a linear mixed model, to account for both fixed effects (such as the specific AI being used) and random effects (such as variability among papers or questions; see S2. Statistical Methods for model specification). We found significant differences in the quality of extractions produced by the AI systems. Elicit performed the best, significantly outperforming both GPT4x1 ($p < 0.001$) and GPT4x3 ($p < 0.001$), indicating that its extractions were closer to the quality considered acceptable by human reviewers. Additionally, GPT4x3 was found to be significantly better than GPT4x1 ($p = 0.006$), suggesting that the iterative approach of running GPT4-Turbo multiple times yielded improved results over a single run.

Manual Review of Flagged Paper-Question Pairs

To ensure the validity of our findings and further understand the discrepancies between AI and human extractions, we conducted a manual investigation of 56 paper-question pairs. These paper-question pairs were flagged by the independent evaluation team during their assessment, and were raised for a variety of reasons, including the need for additional context from the full-text to ensure a fair evaluation, or instances where there was a stark contrast between human and AI responses. Upon re-evaluation, we found that in 28 cases, the AI provided appropriate responses that faithfully represented the content of the study, often with more detail or specificity than the human extractor. This included instances where the human extractor missed certain details, for instance, failing to report all the countries mentioned in a paper, which the AI correctly identified. However, there were also 7 instances where the AI clearly misinterpreted the question, leading to misrepresentations of the study content. The remaining 21 outputs presented a challenge in classification; they contained truthful content that was related to the question but was too general or broad, potentially making them less useful for targeted queries. For example, in one such case, when asked to report the country of the case study, GPT4x1 reported ‘the South Pacific including Solomon Islands’. In another case, Elicit provided correct information about how the community monitors their resources, but also included several other passages of context that were not relevant to the question.

Drivers of AI Extraction Quality

We found that the assessed quality scores were not associated with question content ($p = 0.067$). We noted differences in the quality of responses across different questions, although amongst the factors that we investigated, we did not find any strong evidence to draw conclusions about what drives these differences. Despite there being no overall effect of the question on the assessed quality values, we did find three pairs of questions were significantly different in terms of the evaluated qualities. These were Q1 (‘Which country was the study conducted in?’) vs. Q2 (‘Provide some background as to the drivers and/or motivators of community-based fisheries management.’) ($p < 0.001$), Q1 vs. Q7 (‘How was the data on benefits collected?’) ($p < 0.001$), and Q3 (‘What management mechanisms are used’) vs. Q9

(‘Guidelines for future implementation of CBFM?’) ($p < 0.001$). The rest of the pair-wise evaluations were not statistically different.

We found that the assessed quality of the AI responses was not significantly related to difficulty ($p = 0.055$) of the paper-question pair, as ranked by the human extraction team, nor was there a significant interaction between the difficulty and the AI implementation ($p = 0.877$)(Fig. 4). Therefore, regardless of the difficulty of the paper-question pair, the mean assessed quality was not affected for any of the AI implementations.

3. Discussion

3.1. How well can AI tools discern the presence or absence of relevant contextual data?

These results provide evidence that specialised AI text-extraction tools like Elicit, and general large-language models like GPT4-Turbo, can facilitate the extraction stage of evidence synthesis. Amongst the AI implementations that we tested, Elicit consistently provided higher quality responses compared to our implementations based on GPT4-Turbo, and also had a negligible false negative rate, which means it was much less prone to missing relevant information. We also showed that relying on a generalist LLM like GPT4-Turbo may not be immediately effective, but also that its performance can be improved through aggregating repeated identical calls.

The AI tools that we tested were not reliable enough to be trusted to perform the task without human involvement. While Elicit outperformed the other tools tested, the drawback of its performance was that its responses tended to include a large quantity of unnecessary extra information in addition to the relevant material. In the context of evidence synthesis, this is preferable to failing to deliver relevant information but means that additional human work is still required to ‘sift out’ the relevant material from responses. Compared to other studies which evaluated the effectiveness of LLMs to extract simpler entity or numerical data, we found lower rates of agreement and higher error rates (27, 28).

3.2. How does the quality of AI extraction outputs compare to those done by human researchers?

In the context of actually deploying these tools in a systematic evidence synthesis, our findings suggest that human oversight is still important in this extraction stage of the process. However, we also show that in some cases the AI responses were in fact better than the human responses, either due to level of detail or the ability to find relevant information. We are not familiar with any other studies which have evaluated the ability of LLMs to extract detailed contextual data as we have done here, although we note that automated approaches are being explored (29).

The implications of our findings indicate a future collaborative approach to evidence synthesis with AI tools. By demonstrating that AI can sometimes outperform humans in terms of detail and information retrieval, our research suggests that AI tools have the potential to enhance the breadth and depth of systematic reviews. The prospect of reproducibility inherent in AI-based approaches could also lead to more consistent and verifiable outcomes across different studies. This benefit is particularly relevant given the increasing complexity and volume of research data that needs to be synthesised.

However, the limitations imposed by the proprietary nature of the AI tools used in our study raise concerns about the accessibility and generalizability of these methods. It is essential to continue refining AI tools to ensure that they are transparent and can be adopted broadly by the research community. In doing so, we can work towards a future where AI-assisted evidence synthesis is not only more efficient but also open and reproducible, thus upholding the standards of scientific rigour and integrity.

3.3. Does the difficulty or type of question influence the performance of the extraction?

Question difficulty, as rated subjectively by the human extractors, was not strongly correlated with the quality of AI outputs provides evidence to suggest that AI tools can perform well with documents that might be otherwise difficult for humans to extract information from. However, we note that in the case of a human driven review, there is a learning process that emerges from attempting to apply one's analytical schema to a body of literature, failing, and then adapting, which is a key element of formal methodologies in this space (30–32). This can lead to an iterative refinement of the research questions and goals, which would not be possible under the AI-enabled paradigm that we have tested here. Nevertheless, due to the nature of the AI tools, upon uncovering new considerations or nuances, it could be fairly easy to modify the approach and re-run the extraction, rather than having a team of humans re-read an entire corpus of literature to pick up a previously omitted detail.

This suggests that when undertaking these kinds of systematic reviews, collaborative approaches between humans and AI may prove to be more fruitful than compared to each working in isolation (33, 34). We imagine that there could be a number of ways to improve and develop such a collaborative workflow to improve the quality of evidence syntheses. In line with practice in this study, AI tools could be deployed independent of a human extraction, and both datasets could be combined to produce a synthesis. Adding one or a diversity of AI reviewers in this way could potentially serve to reduce the bias present in a review (13) by providing additional 'perspectives' to those that the human reviewer team has, or as an additional coder to calculate intercoder reliability when undertaking a review alone. Or alternatively, AI tools could be implemented as the first stage of the synthesis and then the human could proof the AI outputs and/or cross-check a sample of data points. In this approach, human reviewers could use the near-zero rate of false negatives (i.e. returning responses to all questions) as a starting point to manually eliminate the false positives.

Future work could explore a range of workflows in a rigorous way to identify promising best practices that prioritise the quality and precision of the overall extraction procedure. We note that as AI tools continue to improve in capabilities, it will be important to continuously benchmark tools - perhaps automatically (29). And whilst general benchmarks for text extraction exist (15), the nuances of contextual understanding present in the field of environmental management, and the importance of fully reckoning with the complexities of environmental systems, will likely require domain specific benchmarks.

3.4. Limitations and future work

Our study suggests that AI tools could be feasibly integrated into evidence syntheses to both accelerate and improve the reliability of the outcomes. However, there are several limitations that must be acknowledged. In this study, due to the capacity constraints of our human research team, we only explored one research topic - CBFM - as a case study for evidence extraction. This allows in-depth insights into some potential strengths and weaknesses of AI in evidence extraction, but future work will need to be done to verify that these findings hold true across other topics. The suitability of AI tools for evidence synthesis may vary across topics that have stronger components of equity, local and traditional knowledge, or ethical considerations - where the danger of misinterpretation is higher. Whilst we have demonstrated some potential workflows for integrating AI into evidence synthesis, more detailed guidance and standards are needed for researchers to outline best practices. Developing these guidelines will necessitate ongoing experimentation, reflection, and refinement of the collaborative processes between human researchers and AI tools - ensuring that the AI is used as a tool rather than replacement for human expertise.

Methods

Overview

We undertook a parallel data extraction process with a team of human reviewers and a set of AI implementations and then evaluated the quality of the AI responses using an independent evaluation team. Two teams were assembled and each given a set of relevant literature for analysis. The first team consisted of human extractors (KO, JD, DK, RS, RAJ) with experience in various forms of evidence synthesis (e.g., (31, 35)), who undertook a traditional evidence extraction from a set of peer-reviewed literature. The other team (SS, FB, MA, RT) with modelling expertise, employed three AI implementations to produce outputs in accordance with the methods that would be followed by the first team. Both the human reviewers and AI models were asked to answer a series of qualitative questions about each of the CBFM-related papers. The AI-enabled team then compared the human responses to the AI model responses and ranked the level of similarity.

Topic

Given the expertise of the human reviewers and the previous screening research (Spillias et al. 2024), this case study continued investigating and annotating data from the existing literature on community-based fisheries management (CBFM). CBFM is an approach to fisheries management where local coastal communities and fishers are responsible for managing their coastal region and resources in an effort to ensure their sustainable use (36). Given the variability and diversity of terminology and language used in CBFM research, this literature provides a robust test of the capabilities of AI to extract and synthesise relevant information.. We chose this case study as it provided the opportunity to elicit a range of data types and complexities to test AI.

Human extraction procedure

The human reviewers created an initial list of questions. In an initial pilot round, all reviewers analysed three randomly chosen papers. The extracted data was compiled and compared by one reviewer (KO) to point to similarities and differences. Then, the human team met to discuss their analysis, differences, and possible ambiguities in the question phrasing. As a result, the extraction questions were modified to form a final list of eleven questions accompanied by a short explanation for each one (Table 1). After the pilot round, 93 papers found in Spillias et al. (2024) were randomly distributed equally among the five human reviewers. When reading the full-text paper, the reviewers could exclude the paper from further review if they thought that it did not fit the research question and/or if the extraction questions could not be answered based on the content. The reviewers excluded 60 papers this way, leaving 33 papers for the full analysis. For each question and paper, the reviewers gave a short answer to the question (Response) and identified one or more passages that supported the answer (Context). If no answer was possible because the information was not in the paper, no answer was given. Because we were interested in identifying drivers of quality in AI extracted results, the human reviewers also recorded their perceived difficulty in answering the extraction question for each paper as either Easy, Medium, or Hard.

Table 1
Contextual Questions and Coding Notes posed to the Human and AI extractors

Question	*Coding Notes*
Which country was the study conducted in?	*If multiple countries are part of a single case study, or if there are multiple case studies - code them separately at the country level*
Provide some background as to the drivers and/or motivators of community-based fisheries management.	*Why is the CBFM in place/being used? Separate to the benefits of the CBFM - could be because there is strong existing community ownership, could be because all other approaches have failed, could be because of an inherent mistrust of government?*
What management mechanisms are used?	*Tangible mechanisms being used to manage the fishery - could be physical limitations like gear limits, size limits or timing limits.*
Which groups of people are involved in the management as part of the CBFM case-studies? Choices: Community-members, Researchers, Practitioners, Government, NGOs, Other	*Involved directly in the management of the fishery, NOT in the conducting of the study (collecting data for the study)*
What benefits of Community-Based Fisheries Management are reported in this case study?	*Physical or social-ecological benefits - could include things like increased fish counts or larger fish size, or improved social outcomes for communities*
What are the indicators of success of CBFM?	*What are the data sources for measuring success - community perception, fish size etc.*
How was the data on benefits collected?	*What methods were adopted - catch and release programs, qualitative survey etc.*
What are the reported barriers to success of Community-Based Fisheries Management?	*Focus on things that are reported around the case study in question - NOT general examples. What was hindering the success of the CBFM - e.g. poor community buy in, poaching etc.*
Guidelines for future implementation of CBFM?	*This can be more general. In light of the study, what are the take home messages for future CBFM projects - NOT future research directions, or things to consider for future studies*
How does the community monitor the system they are managing?	*Within the case study, what are the community groups you have already identified doing to monitor the fishery - are they conducting fish surveys, monitoring community catches etc*
How does the community make decisions?	*How do they make decisions around the management of the fishery - eg. all decisions are passed on by the matai, or individual villages make decisions on any thing that happens from their shoreline.*

AI extraction procedure

We used three AI implementations to extract data from the 33 papers kept by the human reviewers: (i) One call to GPT-4-Turbo (GPT4x1), (ii) three calls to GPT-4-Turbo and synthesised/summarised by a single call to GPT-4-Turbo (GPT4x3), and (iii) the data extraction feature offered by Elicit (Elicit.com). The scripts used to access GPT-4-Turbo via the API are available online at (https://github.com/s-spillias/AI_Extraction; also see Table 1 for prompts).

We accessed GPT-4-Turbo using the Microsoft Azure API in late January 2024. Mirroring the procedure of the human reviewers, the prompt to GPT-4-Turbo (GPT4x1) explicitly requested that the AI return both Response (answer to the extraction question) and Context (a passage of text from the paper that supports the Response). The AI was also prompted to return the output 'No Data/No Context', if it was not possible to answer the question given the paper. Following the prompt, we provided the AI with a cleaned version of the paper text, with metadata and backmatter removed. For the implementation GPT4x3, we employed the preceding strategy three times, harvested the 'Response' portion of the output, and then synthesised the Responses, again using GPT-4-Turbo (see Table 1 for prompt). The Context passages were automatically concatenated, rather than synthesised or otherwise modified, to ensure the accurate capture of the relevant passages.

To produce extraction data from Elicit we uploaded the papers as pdfs to Elicit's online portal on January 29th 2024. We provided a short column name for each question and then put the unaltered text of the question and explanation into the 'Description' and 'Instructions' fields, respectively (See Fig. 5). We also enabled the feature 'High-accuracy Mode' for all columns. Elicit automatically returns 'Supporting' and 'Reasoning' passages from the text to support the response provided, which we used as 'Context' for our evaluation.

For the context strings returned by the AI, we verified their presence in the articles (i.e., that the AI had not 'hallucinated'), by string matching the context returned by the AI with the full-text of the article. We identified 29 instances where the Context returned by the AI did not match any strings in the full-text. We manually investigated each one and confirmed that, with the exception of one unique Context string, all of the Context strings were present in their respective articles. Indicating essentially near-zero rates of hallucination.

Evaluation procedure

Following the human extraction process, and working independently from the extraction team, an evaluation team (SS, MA, RT, FB) developed a procedure for evaluating the quality of the AI outputs in comparison to the human extractions. Initially, the possibility of a fully blind process was considered, where the source of the extractions (human or AI) would not be revealed to the evaluators. However, this approach was ultimately rejected because it was believed that without knowing the source of the extractions, the evaluation team might lack a proper baseline for judging the quality of the responses. Consequently, while the evaluators were aware of which extractions were AI-generated and which were human-generated, they were not informed about which specific AI model produced each extraction. This partial blinding was intended to reduce bias in the evaluation process while still providing a reference point for quality assessment. The resulting procedure is reflective of other evaluation processes developed by other groups (37)d

To assess the quality of the AI-generated responses, the evaluation team established three criteria: i) whether the Context provided by the AI was appropriate evidence for the question; ii) whether the Response was an appropriate synthesis of the Context in response to the extraction question; and iii) how the AI output compared to the human output, with the human extractions serving as a 'Gold Standard'. A three-point scale— -1 for Poor, 0 for Fair, and 1 for Good—was employed by the evaluation team for grading purposes. A custom program was designed in Python to facilitate this evaluation process (Fig. 6). We also performed the same evaluation procedure using GPT4-Turbo as an evaluator to provide further support for the assessments of the human evaluators. This was done through the API and was repeated five times for each question-paper pair).

During the initial assessment, the evaluation team found that some response pairings required the additional context of the paper to evaluate. Therefore a Flag criterion was added to the custom program for further follow-up to investigate the possibility that AI responses had provided more detailed and/or accurate information than the Gold Standard human extraction.

Follow-up Verification

We performed two manual follow-up verification checks to ensure that the AI was returning valid responses from the articles. First, for every AI output, we verified that the Context returned by the AI was indeed present in the full text article. We did this using an automated script that involved a three-step verification process. This process was designed to account for potential discrepancies due to OCR errors, formatting changes, and minor variations in the text. The first step in the process involved normalising the context string extracted by the AI and the corresponding passage from the full-text article. Normalisation entailed removing all punctuation, converting all characters to lowercase, and replacing newline characters with spaces. This step reduces the variability between strings caused by differences in formatting and case sensitivity. The normalised strings were then compared for similarity. We utilised the fuzz.partial_ratio function from the fuzzywuzzy library to calculate the degree of similarity between the AI-generated context and the text from the full article. If the similarity score exceeded a predefined threshold of 90%, the strings were considered similar enough to be a match. This step allows for a degree of tolerance in the match, accommodating minor differences in wording or spelling that might occur. In addition to the string similarity check, we also performed a cosine similarity calculation using the CountVectorizer from the scikit-learn library (38). The two normalised strings were transformed into a document-term matrix, representing the frequency of terms within each string. The cosine similarity between these two term frequency vectors was then calculated. A cosine similarity score greater than 0.5 was considered indicative of a significant match between the strings. The context string was considered valid if it was contained within the full-text passage, exhibited a high degree of similarity with the passage, or had a cosine similarity score indicating a strong match. By employing these methods, we were able to robustly verify the AI-generated context against the source articles, ensuring that the AI did not 'hallucinate' or fabricate information not present in the original texts.

Second, the evaluation team had the opportunity to ‘flag’ question/output pairs that they felt warranted additional investigation. For each of these flagged data points, two authors (SS, KO) manually investigated the AI response and assessed whether it was faithfully reporting information from the paper.

Statistical Analysis

We calculated the inter-rater reliability by using Cohen’s kappa statistic to evaluate the agreement between the presence/absence of contextual data between the human reviewers and the AI implementations.

We created confusion matrices to compare the frequencies at which each AI implementation and human extractors returned similar responses by either providing a response (data) or not (no data) to a given question from each paper. This was calculated separately for each AI implementation (Elicit, GPT4x1, and GPT4x3) and independently for each response value. All values in each confusion matrix are reported as a portion of the total of 1. These matrices have four quadrants and shows how frequently: the human did not provide a response but the AI did (quadrant 1, top right, false positive), the human and the AI both provided a response (quadrant 2, top left, true positive), the human provided a response but the AI did not (quadrant 3, bottom left, false negative), and neither the human or AI provided a response (quadrant 4, bottom right, true negative).

Several criteria were measured throughout this study including how well the AI pulled the relevant context from the paper for the specific question (Context to Question), how well the AI responded to the relevant context (Response to Context), and how the AI response compared to the human response (AI Response to Human Response).

To analyse all the results, we performed statistical analyses using R software (39). Significance was determined using an α-critical level of 0.05 for all tests. A two-way t-test was done to allow for comparison between assessed values for each AI implementation and each criteria against the ‘fair’ extraction score of 0. In subsequent analyses, we employed a linear mixed-modelling approach (lmer function from *lme4* package (40) to assess the impacts of different factors (AI implementations, difficulty, and questions) on overall assessed values. Where applicable, evaluator, extractor, and paper were included as random effects. An analysis of variance (ANOVA) test was performed to assess the overall significance of each variable. When required, pairwise comparisons were then performed on the estimated marginal means (emmeans and pairs functions from *emmeans* package (41) and Tukey’s method was applied to determine significance of comparison.

To examine the overall results, we created a linear mixed-effects model showing the effect of each AI implementation and criteria on assessed quality (e.g. see S2. Statistical Methods for model specification). This model included the effects of the AI implementations and the interaction of criteria as fixed effects, with evaluator and paper as random variables. Additionally, we constructed individual models to assess the effect of the AI implementation on assessed quality for each criterion independently.

We then investigated the impacts of the questions on the assessed quality of the AI responses using a linear mixed-effects model with the question being the fixed effect and evaluator and paper as random effects.

To evaluate the effect of difficulty, as ranked by the human extraction team, on the assessed quality, we used a linear mixed-effects model including the interaction between difficulty and AI as the fixed effect, with paper and extractor as random effects.. The model investigates the relationship between the assessed quality of each question-paper pair and the interaction between the ranked difficulty of the data point and the type of AI implementation.

Finally, a linear model was also created to ensure that the extractor did not have an effect on the assessed quality.

Declarations

The participants were informed of the study and consented to participate.

Acknowledgements

The authors would like to thank Beth Fulton for comments on a previous version of this manuscript. S.S. was supported by a CSIRO R+ CERC postdoctoral fellowship. K.O. was supported by funding from the Centre for Marine Socioecology.

Data and Code Availability

All of the supporting code and data are available at the GitHub repository: https://github.com/s-spillias/AI_Extraction

Use of Generative AI

During the preparation of this work the authors used GPT4-Turbo through Microsoft Azure in order to proofread the manuscript text for errors and clarity. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare no competing interest.

References

1. World Health Organization. Evidence, policy, impact: WHO guide for evidence-informed decision-making. 2021;
2. Cvitanovic C, Shellock RJ, Mackay M, Van Putten EI, Karcher DB, Dickey-Collas M, et al. Strategies for building and managing 'trust' to enable knowledge exchange at the interface of environmental science and policy. Environ Sci Policy. 2021 Sep;123:179–89.
3. Cooke SJ, Cook CN, Nguyen VM, Walsh JC, Young N, Cvitanovic C, et al. Environmental evidence in action: on the science and practice of evidence synthesis and evidence-based decision-making. Environ Evid. 2023 May 18;12(1):10.
4. Walsh JC, Dicks LV, Raymond CM, Sutherland WJ. A typology of barriers and enablers of scientific evidence use in conservation practice. J Environ Manage. 2019 Nov;250:109481.
5. Haddaway NR, Macura B, Whaley P, Pullin AS. ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. Environ Evid. 2018 Dec;7(1):7.
6. Haddaway NR, Bernes C, Jonsson BG, Hedlund K. The benefits of systematic mapping to evidence-based environmental management. Ambio. 2016 Sep;45(5):613–20.
7. Wyborn C, Louder E, Harrison J, Montambault J, Montana J, Ryan M, et al. Understanding the impacts of research synthesis. Environ Sci Policy. 2018;86:72–84.
8. Cook CN, Hockings M, Carter R. Conservation in the dark? The information used to support management decisions. Front Ecol Environ. 2010;8(4):181–6.
9. Jerrim J, de Vries R. The limitations of quantitative social science for informing public policy. Evid Policy. 2017;13(1):117–33.
10. Pullin AS, Knight TM, Stone DA, Charman K. Do conservation managers use scientific evidence to support their decision-making? Biol Conserv. 2004;119(2):245–52.
11. Mallett R, Hagen-Zanker J, Slater R, Duvendack M. The benefits and challenges of using systematic reviews in international development research. J Dev Eff. 2012;4(3):445–55.
12. Curcic D. Number of Academic Papers Published Per Year – WordsRated [Internet]. 2023 [cited 2024 Jul 23]. Available from: https://wordsrated.com/number-of-academic-papers-published-per-year/
13. Haddaway NR, Bethel A, Dicks LV, Koricheva J, Macura B, Petrokofsky G, et al. Eight problems with literature reviews and how to fix them. Nat Ecol Evol. 2020 Oct 12;4(12):1582–9.
14. da Silva Júnior EM, Dutra ML. A roadmap toward the automatic composition of systematic literature reviews. Iberoam J Sci Meas Commun. 2021;
15. Perrault R, Clark J. Artificial Intelligence Index Report 2024. 2024;
16. Berrang-Ford L, Sietsma AJ, Callaghan M, Minx JC, Scheelbeek PF, Haddaway NR, et al. Systematic mapping of global research on climate and health: a machine learning review. Lancet Planet Health. 2021;5(8):e514–25.
17. De La Torre-López J, Ramírez A, Romero JR. Artificial intelligence to automate the systematic review of scientific literature. Computing [Internet]. 2023 May 11 [cited 2023 May 22]; Available from:

https://link.springer.com/10.1007/s00607-023-01181-x

18. Shaib C, Li ML, Joseph S, Marshall IJ, Li JJ, Wallace BC. Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success). 2023;
19. Spillias S, Tuohy P, Andreotta M, Annand-Jones R, Boschetti F, Cvitanovic C, et al. Human-AI collaboration to identify literature for evidence synthesis. Cell Rep Sustain. 2023;
20. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2021 May;133:140–51.
21. Wagner G, Lukyanenko R, Paré G. Artificial intelligence and the conduct of literature reviews. J Inf Technol. 2022;37(2):209–26.
22. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev. 2015;4:1–16.
23. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019 Dec;8(1):163, s13643-019-1074–9.
24. Schmidt L, Mutlu ANF, Elmore R, Olorisade BK, Thomas J, Higgins JP. Data extraction methods for systematic review (semi) automation: Update of a living systematic review. F1000Research. 2021;10.
25. Bolanos F, Salatino A, Osborne F, Motta E. Artificial intelligence for literature reviews: Opportunities and challenges. ArXiv Prepr ArXiv240208565. 2024;
26. Schmidt L, Hair K, Graziozi S, Campbell F, Kapp C, Khanteymoori A, et al. Exploring the use of a Large Language Model for data extraction in systematic reviews: a rapid feasibility study.
27. Zhang H, Jethani N, Jones S, Genes N, Major VJ, Jaffe IS, et al. Evaluating large language models in extracting cognitive exam dates and scores. medRxiv. 2023;
28. Sun Z, Zhang R, Doi SA, Furuya-Kanamori L, Yu T, Lin L, et al. How good are large language models for automated data extraction from randomized trials? medRxiv. 2024;2024–02.
29. Es S, James J, Espinosa-Anke L, Schockaert S. RAGAS: Automated Evaluation of Retrieval Augmented Generation [Internet]. arXiv; 2023 [cited 2024 Aug 9]. Available from: http://arxiv.org/abs/2309.15217
30. Blythe J, Cvitanovic C. Five organizational features that enable successful interdisciplinary marine research. Front Mar Sci. 2020;7:539111.
31. Duggan J, Cvitanovic C, van Putten I. Measuring sense of place in social-ecological systems: a review of literature and future research needs. Ecosyst People. 2023;19(1):2162968.
32. Norström AV, Cvitanovic C, Löf MF, West S, Wyborn C, Balvanera P, et al. Principles for knowledge co-production in sustainability research. Nat Sustain. 2020;3(3):182–90.
33. Schleiger E, Mason C, Naughtin C, Paris C. Collaborative Intelligence: A scoping review of current applications. Qeios. 2023;

34. Wilson HJ, Daugherty PR. Collaborative intelligence: Humans and AI are joining forces. Harv Bus Rev. 2018;96(4):114–23.
35. Karcher DB, Cvitanovic C, Colvin RM, van Putten IE, Reed MS. Is this what success looks like? Mismatches between the aims, claims, and evidence used to demonstrate impact from knowledge exchange processes at the interface of environmental science and policy. Environ Sci Policy. 2021;125:202–18.
36. Doulman DJ. Community-based fishery management: towards the restoration of traditional practices in the South Pacific. Mar Policy. 1993;17(2):108–17.
37. Doostmohammadi E, Holmström O, Kuhlmann M. How Reliable Are Automatic Evaluation Methods for Instruction-Tuned LLMs?
38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
39. R Core Team R. R: A language and environment for statistical computing. 2013;
40. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. ArXiv Prepr ArXiv14065823. 2014;
41. Lenth R, Lenth MR. Package 'lsmeans'. Am Stat. 2018;34(4):216–21.

Figures

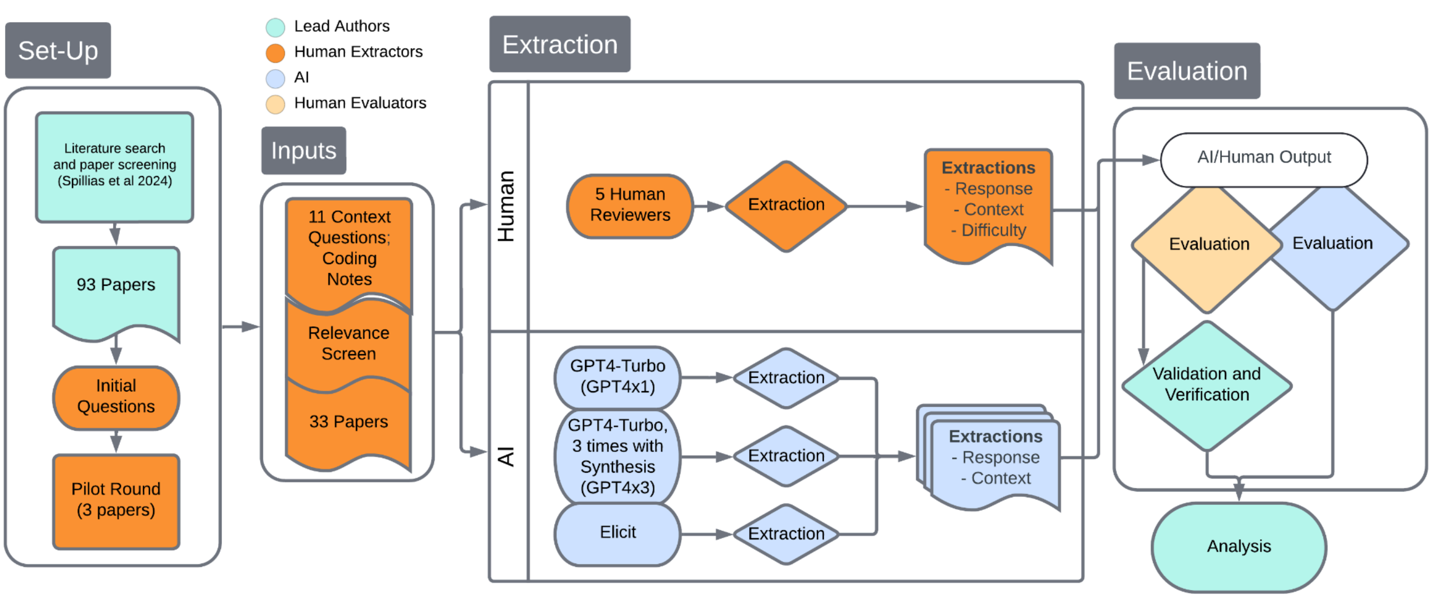

Figure 1

Conceptual Diagram of the Methods. The study employed a dual-extraction and evaluation method, with human reviewers and AI models analysing literature on community-based fisheries management (CBFM) and a separate team of human and AI evaluators assessing the quality of the AI outputs.

Figure 2

Agreement between AI and Human Reviewers on the Presence or Absence of Contextual Data in 363 Paper-Question Pairs. The green quadrants show agreement rates: the top-left quadrant indicates the percentage of pairs where both AI and human reviewers agree that contextual data is present, while the bottom-right quadrant shows agreement on the absence of data. The top-right quadrant highlights false positives, where AI identified data that human reviewers did not, and the bottom-left quadrant indicates false negatives, where human reviewers identified data that AI missed. False negatives are particularly critical as they result in the omission of potentially relevant information, whereas false positives can be subsequently filtered out during evidence synthesis.

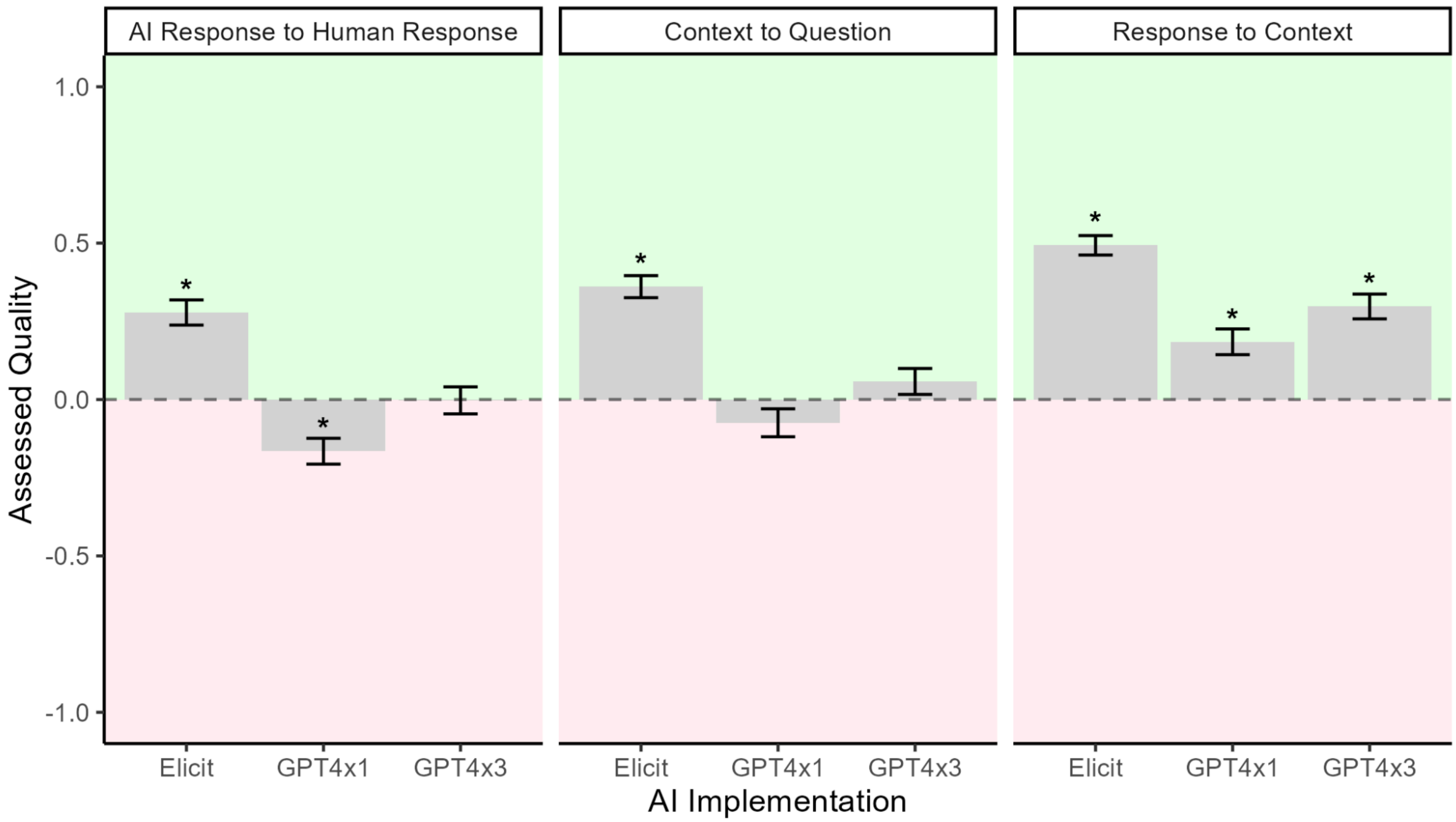

Figure 3

Mean assessed quality of different AI evaluations. The data presented consolidates scores for (i) comparison of AI and human responses, (ii) relevance of context to the question, and (iii) accuracy of the response in reflecting the context. On average, Elicit generally outperformed those of the GPT iterations, indicating a better alignment with the quality standards set by human evaluators. Asterisks represent statistically significant differences from a value of 0.

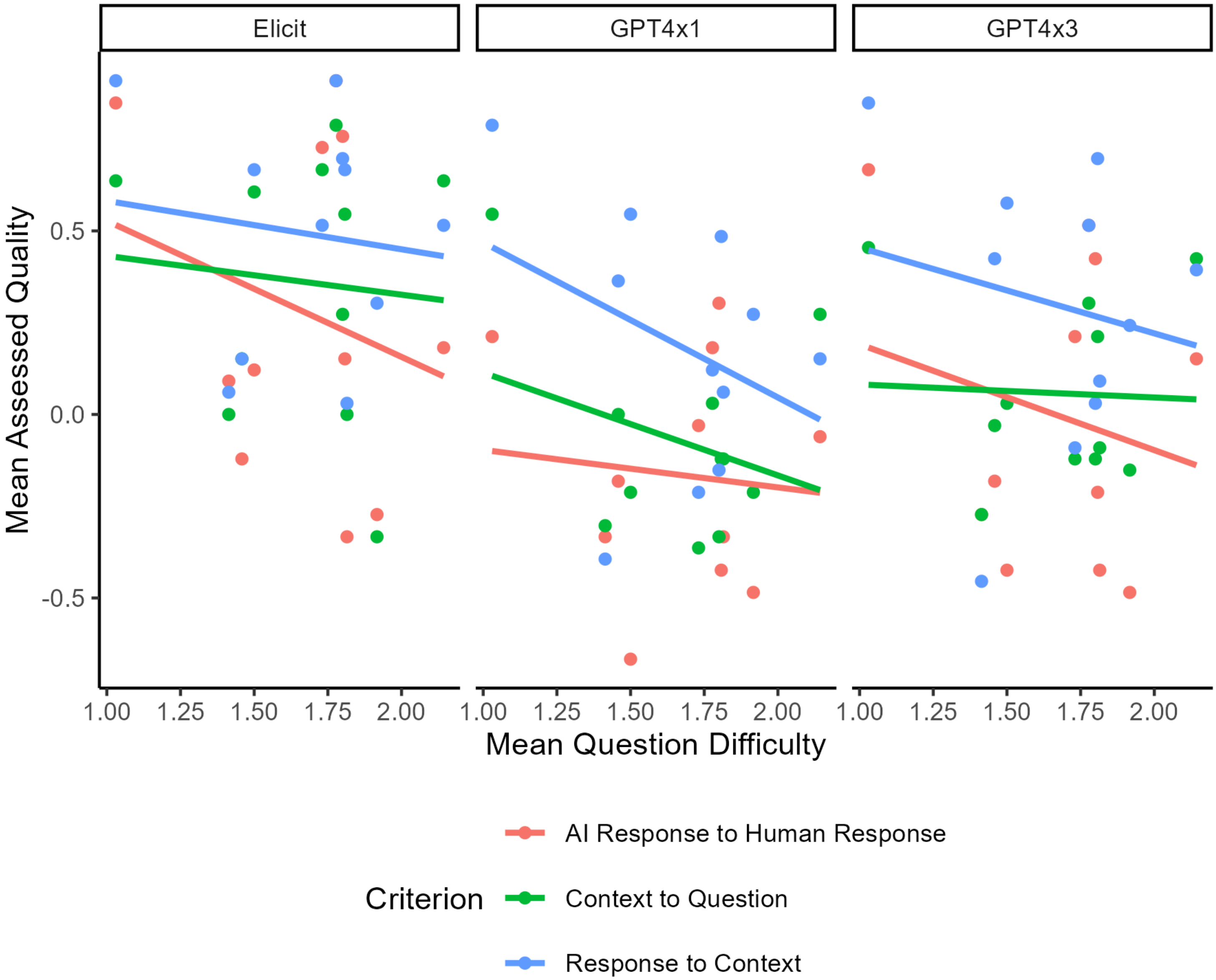

Figure 4

Comparison of AI performance across three implementations (Elicit, GPT4x1, and GPT4x3) in answering questions about scientific literature. The graph plots Mean Assessed Quality against Mean Question Difficulty for three evaluation criteria: AI Response to Human Response, Context to Question, and Response to Context. Each point represents the mean question difficulty across all papers for a specific criterion and question. Trend lines illustrate the relationship between question difficulty and assessed quality for each criterion. The x-axis (Mean Question Difficulty) shows the average difficulty rating assigned by humans to each question, ranging from 1.0 to 2.0. The y-axis (Mean Assessed Quality) indicates how well the AI performed on the tasks, as judged by human evaluators.

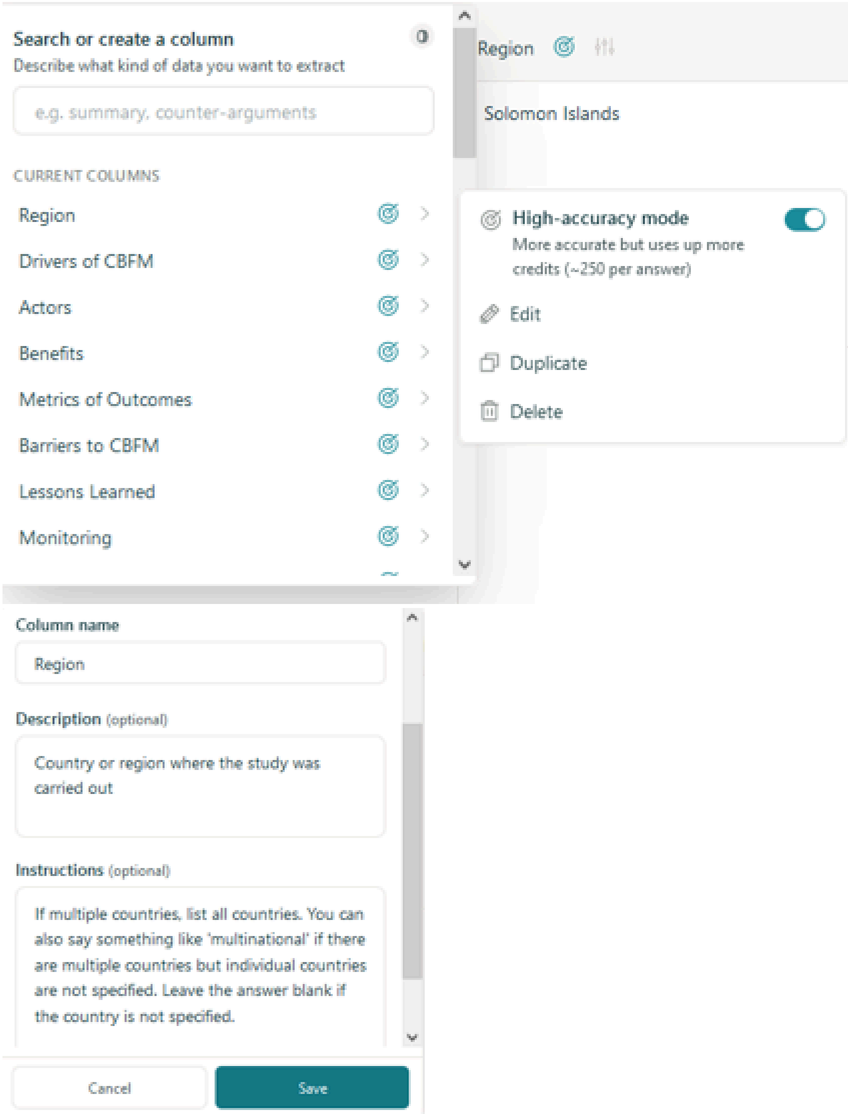

Figure 5

Using Elicit. Screenshot of workflow to demonstrate how extraction questions and coding notes were implemented.

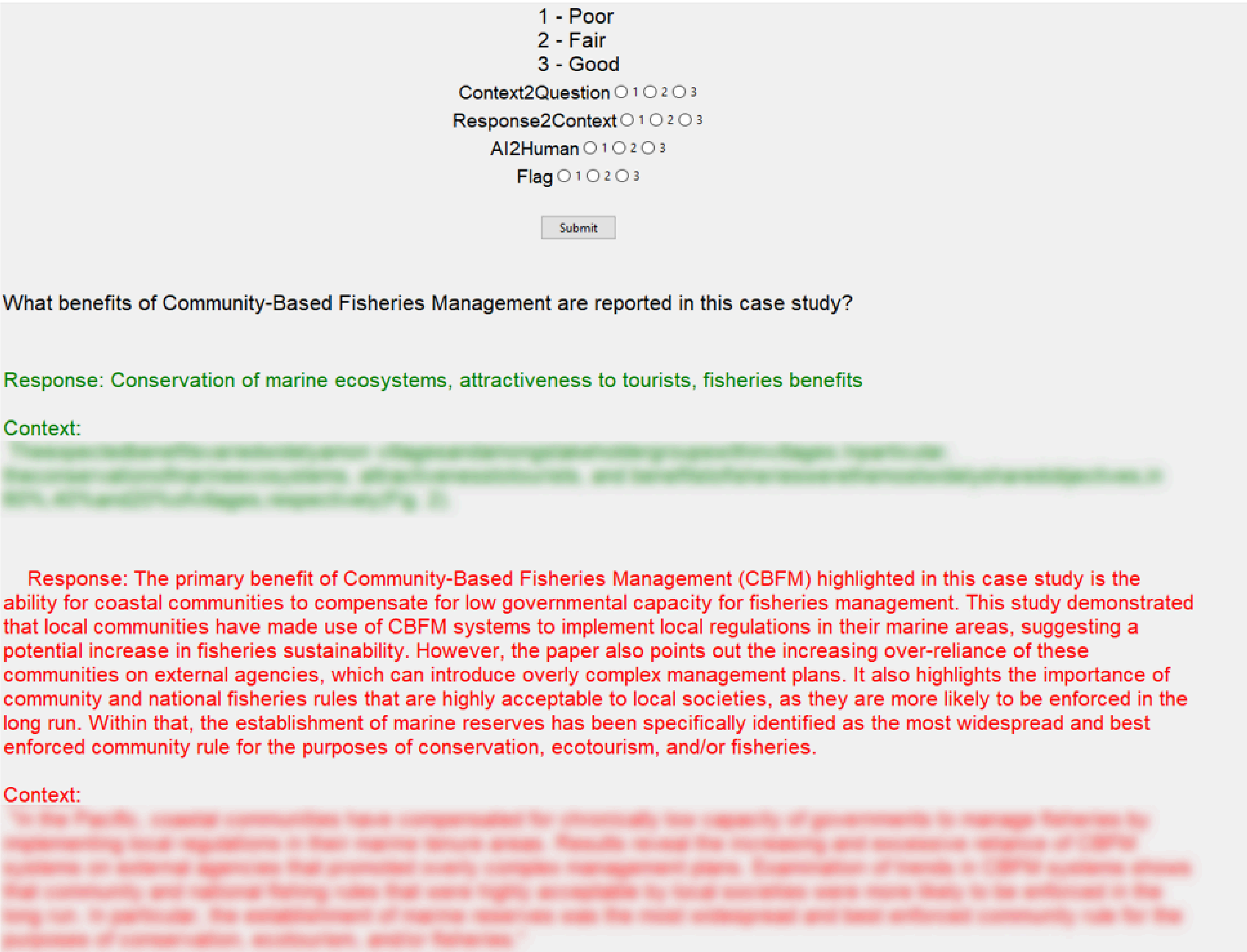

Figure 6

Evaluation procedure. Text in green is from the human reviewers, text in red is from one of the three AI implementations. Specific quotations in the context sections are blurred due to copyright.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplement1.docx